Nucleation of molecular crystals driven by relative information entropy

Gobbo, G., Bellucci, M. A., Tribello, G., & Ciccotti, G. (2017). Nucleation of molecular crystals driven by relative information entropy. *Journal of chemical theory and computation*, 1-14. https://doi.org/10.1021/acs.jctc.7b01027

Published in:
Journal of chemical theory and computation

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright © 2017 American Chemical Society. This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date:05. May. 2023

Nucleation of molecular crystals driven by relative information entropy

Gianpaolo Gobbo,[†,‖] Michael A. Bellucci,[†,‖] Gareth A. Tribello,[‡] Giovanni Ciccotti,[¶,§] and Bernhardt L. Trout[*,†]

†*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

‡*Atomistic Simulation Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, United Kingdom*

¶*Università di Roma La Sapienza, Ple. A. Moro 5, 00185 Roma, Italy*

§*School of Physics, University College of Dublin (UCD), Belfield, Dublin 4, Ireland*

‖*These authors equally contributed to this work*

E-mail: trout@mit.edu

Abstract

Simulating nucleation of molecular crystals is extremely challenging for all but the simplest cases. The challenge lies in formulating effective order parameters that are capable of driving the transition process. In recent years, order parameters based on molecular pair-functions have been successfully used in combination with enhanced sampling techniques to simulate nucleation of simple molecular crystals. However, despite the success of these approaches, we demonstrate that they can fail when applied to more complex cases. In fact, we show that order parameters based on molecular pair-functions, while successful at nucleating benzene, fail for paracetamol. Hence, we introduce a novel approach to formulate order parameters. In our approach, we

construct reduced dimensional distributions of relevant quantities on the fly and then quantify the difference between these distributions and selected reference distributions. By computing the distribution of different quantities and by choosing different reference distributions, it is possible to systematically construct an effective set of order parameters. We then show that our new order parameters are capable of driving the nucleation of ordered states and, in particular, the Form I crystal of paracetamol.

1 Introduction

Crystallization plays an important role in many industrial processes ranging from food production[1] to the preparation of pharmaceutical drugs.[2–4] It is thus unfortunate that our understanding of the earliest stages of crystallization, the so called nucleation stage, is incomplete. This lack of understanding is perhaps not surprising, however, as studying nucleation is challenging both from the experimental and the computational point of view. The experimental study of nucleation is difficult because the onset of nucleation involves exceedingly small time and length scales. As a result, any experimental technique needs to meet stringent resolution criteria to be applicable, which makes the direct experimental characterization of nucleation extremely difficult.[5–7] One might be tempted to think that computer simulations with full atomistic potentials are the ideal tool to provide insight into the mechanism of nucleation. Unfortunately, however, the typical time scale for the occurrence of a random fluctuation that leads to a nucleation event can easily be on the order of hours in realistic conditions, which is several orders of magnitude larger than the time scales that are accessible in molecular dynamics (MD) simulations. Consequently, a simple brute force approach to study nucleation is unfeasible, and enhanced sampling techniques[8–26] are a necessity. Most of these techniques use low-dimensional descriptors of the state of the system called order parameters (OPs), or collective variables. While enhanced sampling techniques are incredibly useful from a simulation standpoint, in practice it can often be difficult to find a set of OPs that are able to distinguish between the various metastable states of the

system of interest, and the transition state regions connecting them, without a detailed prior understanding of where in phase space these metastable regions and transition states lie.

The formulation of OPs for the nucleation of molecular crystals is particularly difficult, and it has only recently become possible to perform simulation studies of nucleation from the melt[27–31] and from solution[32,33] for relatively simple molecular crystals. In these studies, OPs based on parametrized models of molecular pair-distribution functions were used. These pair-function based OPs classify individual molecules as belonging to a crystal state if the distances and/or relative orientations between them and their neighbors are commensurate with those in a target crystal form of interest.[34,35] In this way, when many molecular pairs have relative distances and/or orientations that are characteristic of the crystal, the system is considered to be in a crystal-like state. These OPs have been successfully applied in studies of simple molecular crystals, however, as we show in a paradigmatic case, they can fail for systems of higher complexity. To address this issue, we introduce a novel approach for the construction of OPs for the nucleation of molecular crystals, which is based on comparing distributions. Since any ordered state is characterized by the emergence of long range order correlations that are themselves manifested through the presence of peaks in the distribution of selected structural quantities, we define new OPs by measuring the difference between the instantaneous distribution of selected structural quantities of the system and well chosen reference distributions. Using our approach, it is not only possible to quantify the "distance" between the instantaneous distribution of the system and the distribution of a specific crystal form of interest, but also to formulate OPs that distinguish between an ordered state and a disordered, or less ordered, state. Our approach is systematic in that it allows one to construct OPs of increasing complexity and to include multiple distributions of relevant structural quantities in the description.

The rest of the paper is organized as follows. We first investigate whether the OPs that have been used thus far for studying nucleation in molecular crystals can be used with metadynamics to drive benzene and paracetamol to nucleate from the melt. In particular, we

demonstrate for paracetamol that when these OPs are used to accelerate the sampling, no ordered configurations are visited. We then show how our new approach can be used to construct systematically a set of OPs. We start by constructing an OP aimed at distinguishing ordered arrangements of molecular centers from disordered ones. When this OP is incorporated into metadynamics simulations together with one of the OPs based on pair-functions, there is a dramatic increase in the efficiency of the exploration of phase space, and the system thus visits numerous metastable ordered states that were previously inaccessible. Finally, we consider different structural quantities of the system, and we show how more complex OPs can be built and refined using our approach, and how these OPs can drive nucleation of the Form I crystal of paracetamol.

For the sake of clarity, we also note that the purpose of the simulations discussed in this paper is to demonstrate that OPs constructed using our approach are effective at driving nucleation in a very challenging case. We do not aim to provide detailed insight into the nucleation mechanism in realistic conditions as this will be the subject of future work.

2 Enhanced sampling simulations using pair-function based order parameters

The configuration of a system of small molecules is defined by the complete set of positions of all the atoms of every molecule. However, it can be reduced and simplified through the introduction of what is called a point molecule representation.[34] This consists of a center for each molecule, which can be for instance its center of mass or the position of one of its atoms, and a set of one or more molecule-centered vectors that accounts for the orientation of the molecule in space[†] (see Fig. 1).

When characterizing a molecular crystal structure, not only are the positions of the

[†]If needed, a set of internal degrees of freedom accounting for the internal structure of the molecule can also be defined.

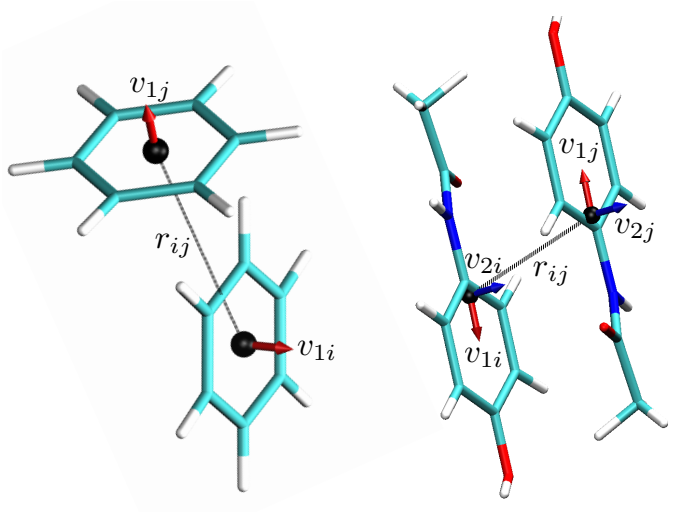

Figure 1: Point molecule representation for benzene, on the left, and paracetamol, on the right. The centers of the molecules and the vectors accounting for the orientation of molecules in space are shown as black dots and colored arrows respectively. Due to molecular symmetry only one vector per molecule is defined for benzene while two vectors define the orientation of a paracetamol molecule. The distance vector separating the center of molecule i from that of molecule j is labeled r_{ij}.

centers of the molecules important, but also their orientations. In particular, the underlying periodicity and order of a molecular crystal ensures that a crystal form is characterized by a specific set of values of relative distances and angles between the molecules. Pair-distribution density functions provide a natural framework to characterize this property as is highlighted in Fig. 2. The first and second panel show the probability density of relative distances and angles between pairs of benzene molecules in the liquid and crystal phase, respectively. These joint distributions were computed from MD simulations at 250 K. Notice that the distribution for the liquid state is quite broad while the distribution for the crystal state displays sharp peaks.

The properties of molecular pair-distributions, or more generally of molecular pair-functions, imply that they can be easily used to construct OPs that try to classify ordered states. In fact the OPs that have been used with enhanced sampling techniques to study the nucleation of molecular crystals identify crystalline states based on an analysis, through pair-

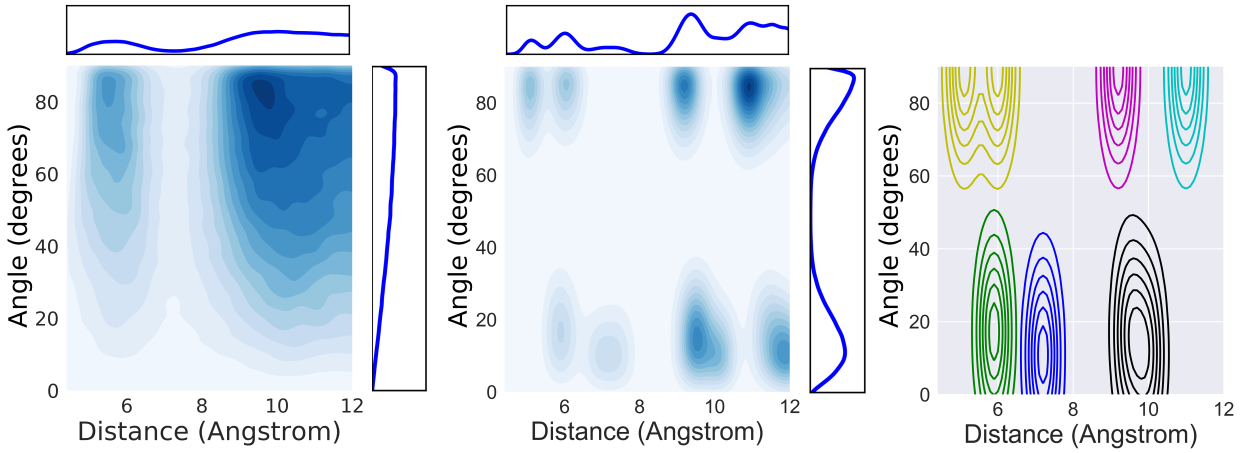

Figure 2: Probability density of distances and relative angles between benzene molecules in the liquid (first panel) and Form I crystal (second panel) state. Due to molecular symmetry the relative angle between vectors ranges from 0 to 90 degrees. The distribution for the liquid state is quite broad while that for the crystal displays narrow peaks. The third panel shows a sum of Gaussians that approximately reproduces the position of the peaks in the second panel.

functions, of the relative distances and/or orientations between neighboring molecules.[34,35] For clarity, we use the term pair function to refer to a localized function that is substantially different from zero only on a compact domain. Simple examples are a Gaussian function or a sum of Gaussians. Santiso and Trout[34] introduced a systematic method for developing these pair-distribution function based OPs. Their approach allows one to define per-molecule OPs that account for the degree of crystallinity in the surroundings of each molecule. These OPs can then be averaged over the whole system, or over portions of it, to build global OPs.

A number of approaches that are similar to Santiso and Trout's have since been developed and we would refer the interested reader to a thorough discussion in Appendix A. Here we will only discuss the specific expressions that we have used in this work. In the case that a point molecule representation is characterized by the position of its molecular center and only one orientation vector, the per-molecule OP takes the form

$$\Gamma_i^{rv} = \frac{1}{n_i} \sum_{j \neq i} s(|r_i - r_j|) \sum_{\alpha=1}^{M} e^{-((|r_i - r_j| - d_\alpha)^2 / 2\sigma_{d_\alpha}^2)} e^{-((\theta(v_i, v_j) - \theta_\alpha)^2 / 2\sigma_{\theta_\alpha}^2)}, \quad (1)$$

where r_i and v_i are the position of the center of molecule i and the vector representing its orientation, respectively. In the equation above, M is the number of peaks in the joint distribution of distances and angles that are being considered. The center of every peak is specified by a parameter for every attribute, *e.g.*, θ_α for one relative angle and d_α for the modulus of the distance, while σ is a free parameter that defines the width of the Gaussian that is used to represent a peak. In the outermost summation in eq. (1), j runs over all the molecules different from i and s is a smooth switching function that selects only the pairs of molecules that are within a certain distance cutoff of each other. In addition, $n_i = \sum_{j \neq i} s(|r_i - r_j|)$ is a smoothed version of the coordination number for molecule i that acts as a normalization factor. A possible choice for the explicit form of the function s is

$$s(r) = \frac{1 - \left(\frac{r}{r_0}\right)^n}{1 - \left(\frac{r}{r_0}\right)^m}, \tag{2}$$

where r_0, n and m are free parameters. The third panel of Fig. 2 shows how this works in practice. Two dimensional unnormalized Gaussians, such as those in the α sum of eq. (1), are used to reproduce the positions of the peaks in the distribution that is shown in the second panel. The rationale behind the construction of these OPs is that a molecule in the crystal phase will typically have relative distances and orientations with most of its neighboring molecules in proximity to one of the M peaks in the reference distribution, and thus the sum over all the molecules will yield a high value. By contrast, a molecule in the liquid phase will have its neighbors distributed more randomly. In particular, the majority will have values for the relative distance and orientations that are away from the peaks in the reference distribution. The summation in the OP will thus give a lower value when it is computed for molecules in the liquid phase. In the case that the point molecule

representation is defined using two vectors, the per-molecule OP easily generalizes to

$$\Gamma_i^{rv_1v_2} = \frac{1}{n_i}\sum_{j\neq i} s(|r_i - r_j|)\sum_{\alpha=1}^{M} \; e^{-((|r_i-r_j|-d_\alpha)^2/2\sigma_{d_\alpha}^2)} e^{-((\theta(v_{1i},v_{1j})-\theta_{1\alpha})^2/2\sigma_{1\alpha}^2)} \\ e^{-((\theta(v_{2i},v_{2j})-\theta_{2\alpha})^2/2\sigma_{2\alpha}^2)}. \tag{3}$$

Enhanced sampling techniques such as metadynamics[14] or TAMD/d-AFED[18,19] that use OPs to drive continuously the full configuration of the system from one metastable state to another, have the advantage of highlighting deficiencies of the OPs. A good set of OPs ensures reversible transitions between the metastable states with clear and sufficient separation between the states. High hysteresis, overlap between the metastable states, or worse failure to drive the system to the desired states are hallmarks of a non-adequate set of OPs. Since all of these potential problems become apparent when using these kinds of simulation techniques, we have decided to use one of these techniques (metadynamics) to test rigorously the capabilities of the OPs considered in this paper.

We have considered, as test cases, simulations of nucleation from the melt in two small systems, 144 benzene molecules and 96 paracetamol molecules. Benzene is a small non-polar molecule, and its nucleation process has been already investigated using different methods.[27–29] Paracetamol is a molecule of great relevance for the pharmaceutical industry. Successful simulation of nucleation of paracetamol from solution would pave the way for a rational in-silico approach to the improvement of the industrial crystallization process and would allow one to investigate heterogeneous nucleation and epitaxy under various conditions.[36–42] However, even simulation of nucleation from the melt is yet to be accomplished. In fact, the complexity of the crystal structure, the strong polar character of the molecule and the high viscosity of the liquid makes the simulation of nucleation extremely challenging.

The paracetamol molecule is composed of an aromatic ring with an OH group attached to one of the carbons and a short 8-atom tail attached to the carbon opposite to the OH group (see Fig. 1). The center of the point molecule representation was thus set so that it

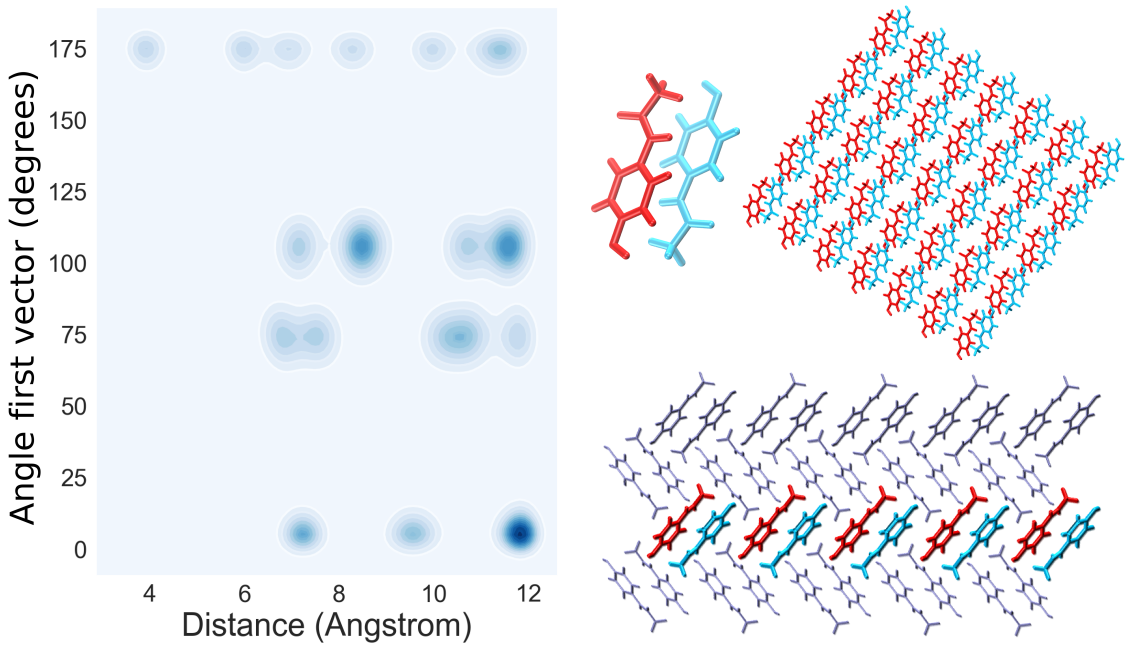

Figure 3: The first panel shows the probability density for the modulus of the distance and the relative angle between the v_1 vectors that was obtained from an MD simulation of paracetamol in crystal form 1. The second panel shows a pictorial representation of the mutual arrangement of paracetamol molecules in the perfect Form I crystal at different levels of detail. Two molecules in an arrangement corresponding to the shortest distance peak (around 4 Å and 180 degrees) in the probability density of the first panel are shown in red and blue. Their disposition in a particular slice of the crystal parallel to the (100) plane (top) and in the full crystal structure (bottom) viewed orthogonal to the (001) plane are also shown.

coincides with the center of mass of the molecule, while the first vector, v_1, was chosen to be the vector that connects the OH carbon to the center of mass. The second vector, v_2, was then set to be orthogonal to the plane defined by the atoms in the carbon ring.

The most stable form of paracetamol at room temperature and pressure conditions is crystal Form I, which is characterized by a monoclinic unit cell. The joint probability distribution for the intermolecular distances and relative angles between the first vectors is shown in the first panel of Fig. 3. This distribution was computed using configurations that were taken from a MD simulation of the Form I crystal at 298 K. Some peculiarities of the crystal structure can be noticed from this distribution. Four different relative orientations between the v_1 vectors of the molecules are possible, but only one occurs at very short

distances. This is because every molecule has its nearest neighbor at a distance of about 3.9 Å, and within these pairs the first vectors, v_1, in our point molecule representations are oriented in an antiparallel fashion so that the whole pair forms a sort of dimeric entity. The second panel of Fig. 3 shows this dimeric subunit and highlights its presence in the crystal structure. In addition, in comparing the joint distribution of paracetamol in Fig. 3 with that of the joint distribution of the crystal state of benzene in Fig. 2, we see that the paracetamol crystal has significantly more peaks, which suggests that it has a much more complex crystal structure than benzene.

In Fig. 4, we show the results of a metadynamics simulation of benzene that was biased using the global average of the OPs defined in eq. (1), *i.e.* $\Gamma^{rv} = \frac{1}{N}\sum_i \Gamma_i^{rv}$, with $N = 144$ being the total number of molecules. From Fig. 4, it is evident that the system undergoes reversible transformations between various metastable states. Direct inspection of the trajectory reveals that some of the metastable states, marked C_1 in the picture, are characterized by a molecular arrangement typical of Form I even though the values of the OP are lower than the reference OP value for the crystal state (dashed green line in Fig. 4) calculated from the unbiased MD simulation. This discrepancy is due to the presence of a few defects as can be seen in the upper right panel of Fig. 4, where a configuration visited around time $t = 68$ ns is shown. A different ordered state, marked C_2 and characterized by a more planar relative orientation between the molecules, is also visited during the simulation. Even though the system visits the crystal and liquid states multiple times, the substantial amount of overlap between the values taken by the OP in the liquid (marked L) and crystalline states indicates that the OP cannot completely resolve these states.

In Fig. 5, we show the results of the application of a pair-function based OP approach to the paracetamol system. The simulation was run biasing two OPs with metadynamics. We used the global average of OPs of the form of equation (3), $\Gamma^{rv_1v_2} = \frac{1}{N}\sum_i \Gamma_i^{rv_1v_2}$, defined with different ranges and cut-off distances (see caption of Fig. 5 for details). Both the OPs get to values that are compatible with those in the crystal state (shown as dashed lines) at

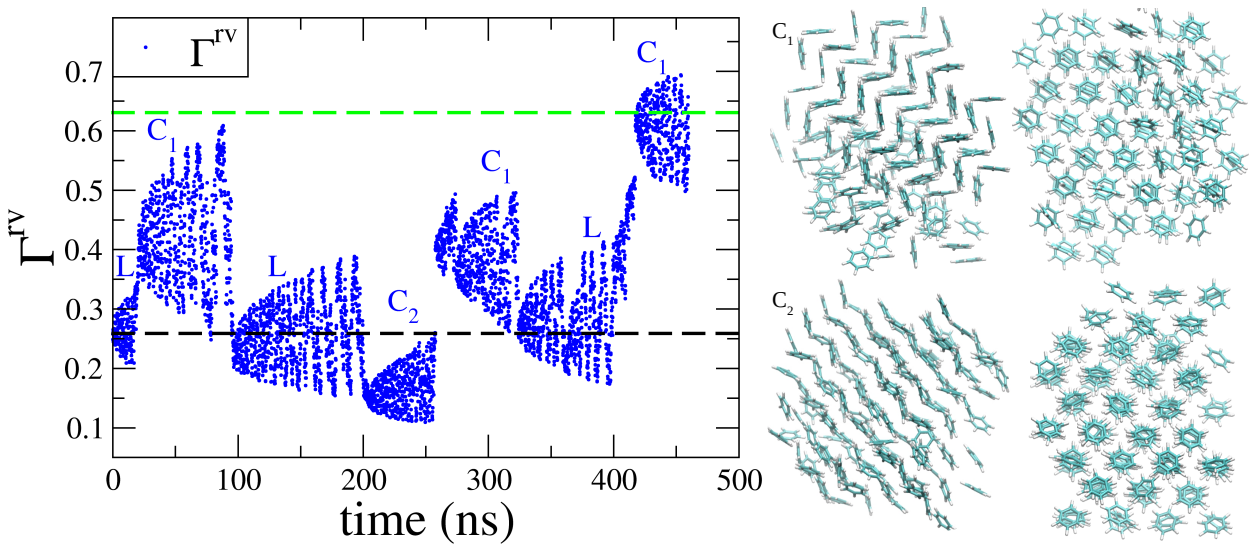

Figure 4: Results from a metadynamics simulation of benzene biasing the molecular pair-function based OP considering both distances and angles, $\Gamma^{rv} = \frac{1}{N}\sum_i \Gamma_i^{rv}$. The Gaussians defining the peaks are those shown in the third panel of Fig. 2. The σ_{θ_α} and σ_{d_α} parameters were chosen to be about 17 degrees (0.3 radians) and 0.3 Å for every α. We used a switching function s of the form of eq. (2) with parameters $r_0 = 10$ Å, $n = 10$ and $m = 20$. The metadynamics bias potential was constructed by deposing Gaussians every 8 ps with a height of 2 kJ/mol and a σ of 0.001. The evolution of the OP in time is shown on the left. The green (black) dashed line represents the typical value of the OP in the crystal (liquid) state. Labels mark the different metastable states visited during the simulation. On the right, two different projected views of a configuration marked as C_1 and one as C_2 are shown. The C_1 configuration was visited around time $t = 68$ ns. The molecular arrangement is that typical of Form I. The presence of defects causes the value of the OP to be lower than the reference value for the crystal (green dashed line). The C_2 configuration was visited around time $t = 256$ ns and is characterized by a disposition of molecules different from that of Form I.

the same time. However, unlike our simulations of benzene, no relevant metastable state or ordered minimum is ever visited. This is a clear signal of the fact that the two OPs are not able to describe the transition and are not even refined enough to define the target crystal state unequivocally. In Appendix B we show some examples of configurations of both systems sampled during the simulations that are misclassified by the OPs and comment on the reason why such misclassifications occur. Finally, we note that, for both the systems considered, we have also verified (data not shown), that varying the values of the parameters and the specific set of attributes involved in the definition of the OPs does not qualitatively

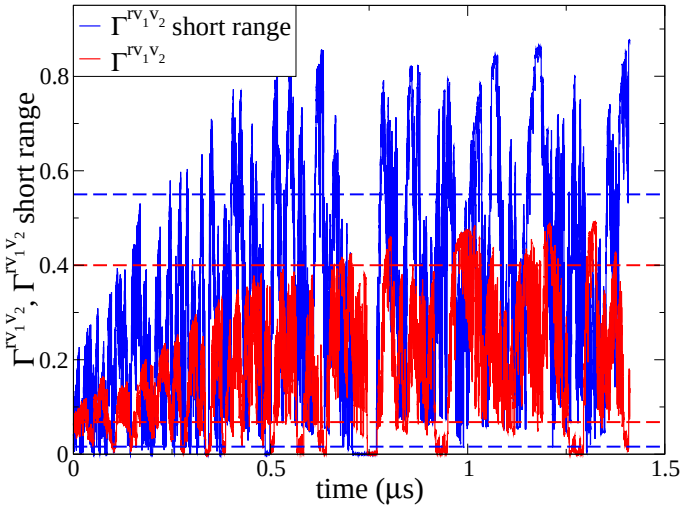

Figure 5: Metadynamics simulation of the paracetamol system. Pair-function based OPs that consider the occurrence of both distances and angles for both v_1 and v_2 vectors at the same time were used. The red series represents the evolution of $\Gamma^{rv_1v_2}$. This OP uses all the peaks in the joint probability density up to 12 Å. The two dashed red lines represent the reference values for the crystal (higher value) and liquid (lower value) state. The second OP is of the same kind of the first but uses only the very first peak in the joint density corresponding to the dimeric entity discussed earlier and characterized by a distance of about 4 Å. Its evolution is shown in blue and the dashed lines represent reference values for the crystal and the liquid state. The distance switching function used were of the form of eq. (2) with parameters $n = 10$, $m = 20$. We set r_0 to 11.5 and 5 Å in the two cases. The metadynamics bias potential was constructed by deposing Gaussians of height 6 kJ/mol every 8 ps and using a width of 0.026 for both OPs.

change the behavior of the simulations.

3 Order parameters based on relative information entropy between distributions

The OPs described above are indicator functions used to characterize the degree of crystallinity of a system. They have the advantage of being local so that a degree of crystallinity can be assigned to each molecule and not just to the system as a whole. However as discussed above, since they are parametrized on a specific crystal polymorph, they are sensitive only in the vicinity of this crystal state. There is no guarantee that other crystal polymorphs can

be distinguished from the liquid state nor between each other. Another drawback is that because these OPs are based on simple pair-functions they are prone to degeneracies. In other words, very different configurations can give similar values for the OP, which in turn, leads to misclassification of the sampled configurations. For example, there is in fact no guarantee that a large value of the pair-function based OP actually corresponds to a configuration that is the target crystal of interest.

We present here a new approach that is based on the analysis of the distributional properties of a system. It is in fact the distribution of certain structural quantities, rather than the occurrence of their single specific values, that classifies a particular ordered state. In other words, it is the emergence of long-range correlations that discriminates an ordered state from a disordered state. In our approach, to construct an OP we select one or more relevant quantities and we build the relative probability density p on the fly. This instantaneous probability distribution is then compared with a suitable chosen reference distribution q.

A natural way of comparing probability distributions, that is also particularly appealing from the physical point of view, is provided by information theory and amounts to computing the relative entropy, also known as Kullback-Leibler divergence[43](KLD), from the probability distribution p to the distribution q:

$$KL(q||p) = \int q(x) \ln \left(\frac{q(x)}{p(x)} \right) dx. \tag{4}$$

The KLD is always non-negative and is zero if and only if $q(x) = p(x)$, $\forall\ x$. In fact, it is often referred to as the distance between two probability densities, even though, being non-symmetric, it is not strictly a distance metric. To construct a differentiable OP starting from eq. (4) one needs first to build smooth probability densities p and q. This can be done through the use of a kernel density estimate[44](KDE). In practice, instead of binning the data as would be normally done with a histogram, one associates to every data point a localized kernel function that is centered on the data point itself. Here, we use Gaussian kernels and

approximate the probability density as

$$p(x) \approx \frac{\sum_i w(x_i) g(|x - x_i|)}{\sum_j w(x_j)}, \tag{5}$$

where the sum runs over the elements of the data set, g is a normalized Gaussian function in one or more dimensions and the weights, w, of the data points may be non-identical. The sum in eq. (5) then gives the desired smooth probability density estimate, the derivatives of which are well defined with respect to any of the data points. To avoid numerical instabilities that can arise when p is very small, a typical numerical workaround[45] can be used: p is regularized with respect to q, *i.e.* $KL(q||(\frac{1}{2}p + \frac{1}{2}q))$ is computed instead of $KL(q||p)$. This quantity is still zero if and only if $p = q$. Finally, to compute the integral numerically, the KDE must be evaluated on a grid. Hence, the OP we compute takes the final form

$$\sum_m q\left(m\right) \ln\left(\frac{q\left(m\right)}{\left(\left(p\left(m\right) + q\left(m\right)\right)/2\right)}\right) dS, \tag{6}$$

where m runs over the points of the grid and dS is the measure of the volume element associated to every grid point.

The idea of comparing distributions using the KLD has been previously used for the inverse design of interactions in colloidal systems.[46] In addition, Gimondi and Salvalaglio[47] have recently used a similar approach to compare unidimensional distributions of relative orientations between molecules as an analysis tool to detect ordered states in nanopore confined simulations of carbon dioxide. However, to the best of our knowledge none of these approaches have been used to drive the sampling in any study of molecular crystal nucleation.

3.1 An order parameter to distinguish order form disorder

The formulation for the construction of OPs just introduced and based on the KLD ensures high flexibility. In fact, one can choose the distribution of any set of quantities and there is ample freedom in the choice of the reference distribution q. Furthermore, one can limit

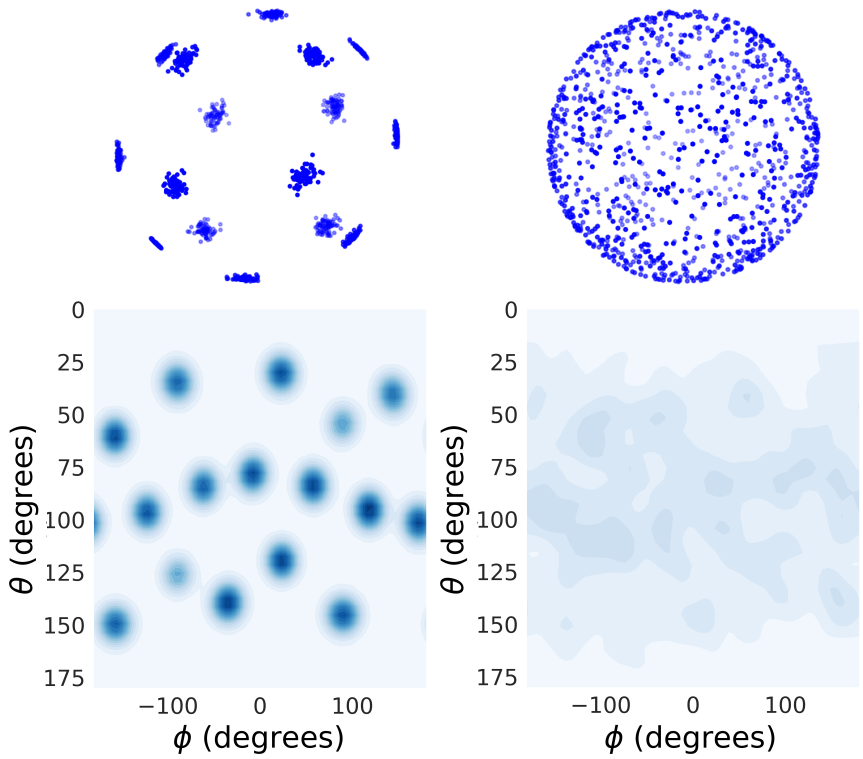

Figure 6: The top left and right panels show the intermolecular distance vectors of magnitude smaller than 8 Å, suitably normalized, for a typical crystal and liquid configuration of benzene, respectively. The lower panels show the corresponding KDE probability density estimate in the space of azimuthal and dihedral angles computed using kernels as described in the main text. The contour levels and the color code for the two plots are the same.

the calculation of the distribution to a subset of molecules or to a particular spatial portion of the system. Quantifying how much the instantaneous distribution of some structural quantities of a system differs from any chosen reference leads naturally and directly to the formulation of OPs aimed at distinguishing a generally ordered phase from a disordered one. In fact, while it is obviously possible to use as reference the distribution for a particular crystal form of interest, it is also possible to use a uniform distribution, or the distribution of the liquid state. Such an OP would be extremely difficult, if not impossible, to formulate using the approach based on pair-functions. Furthermore, the transition from a disordered to an ordered phase is, by definition, characterized by a change in the entropy of the system, *i.e.* by the emergence, and subsequent concentration, of peaks in the distribution of some of the relevant properties of the system, not by their specific location.

Very recently, the idea of using entropy as an OP has been used, alongside enthalpy, to study the nucleation of monoatomic crystals.[48,49] In particular, the authors introduced an OP that builds on the radial distribution function and is an approximated version of the excess entropy per atom. OPs of this kind are of particular interest as they don't make any specific assumption about the nucleation pathway. Starting from the same motivations and following the considerations outlined at the beginning of this section, our approach can also be used to formulate OPs that distinguish order from disorder. In our case, however, we are able to build OPs that use distributions based on a large variety of structural quantities, rather than just the modulus of the distances between atoms, as is the case with the radial distribution function. Therefore, our approach allows one to systematically construct OPs of increasing complexity which is a necessary requirement when treating molecular systems. We now illustrate, as an example, how to construct an OP that accounts for what we term positional ordering, *i.e.* the degree of order in the arrangement of the centers of the molecules in space. We will label the resulting OP $KL^{\hat{r}}$. Starting from the point molecule representation, we consider the set of distance vectors between the centers of the molecules $\{r_i - r_j \mid i,j \in 1,\dots,N; i \neq j\}$, where N is the total number of molecules. Moreover, we limit ourselves only to distance vectors between neighboring molecules. This can be achieved, preserving differentiability, using a smooth switching function for the weights (w in eq. (5)). For this purpose we use a switching function of the form in eq. (2), with $r_0 = 7$ Å. At this point, instead of constructing a weighted KDE for the vectors distribution, we consider the set of normalized distance vectors $\hat{r}_{ij} = (r_i - r_j)/|r_i - r_j|$, *i.e.* points on the unit sphere. For every normalized distance vector we compute the azimuthal angle θ of $\hat{r}_{ij}$ with the z-axis, and the dihedral angle ϕ between the $z\hat{r}_{ij}$-plane and the xz-plane. We then compute the KDE of the distribution of the two angles on a two dimensional grid with domain $[0, 180] \times [-180, 180]$ using Gaussian kernels of width 5.73 degrees (0.1 rads) and 11.46 degrees (0.2 rads) for θ and ϕ, respectively. The ϕ angle is not well defined for $\theta = 0$ or 180, and thus its derivatives can become numerically problematic when the value of θ

is close to one of the poles. To regularize the behavior of the derivatives of ϕ we therefore multiply it by a switching function that acts on θ. This switching function has a value of 1 when θ is more than 10 degrees from one of the poles, *i.e.*, when $10 \leq \theta \leq 170$, and a value of zero when θ is within 5 degrees from the poles. Between these limits the value of the switching function varies smoothly from 0 to 1.

The value of $KL^{\hat{r}}$ is computed according to eq. (6) using the uniform distribution as the reference distribution, q. In general, one could also use the distribution for the liquid state as the reference distribution. However, in the event that this distribution is not particularly structured, using a uniform reference works as well. We also note that, by substituting the set of normalized distance vectors between the centers with the orientation vectors, it is possible to build an OP, KL^{v}, that accounts for the degree of order in the orientations of the molecules.

The upper panels of Fig. 6 show the intermolecular distance vectors that are smaller than 8 Å, suitably normalized, for a specific configuration of the benzene system in the crystal and liquid phase, respectively. These distributions on the unit sphere clearly reveal the order in the crystal phase and the disorder in the liquid phase. The lower panels of Fig. 6 show the corresponding KDE probability density estimate in θ and ϕ. In Fig. 7, we show the value of $KL^{\hat{r}}$ computed from the trajectories of the benzene and paracetamol simulations in which the pair function based OPs were biased and that were discussed in the previous section. For the benzene simulation, it can be seen that there is good separation between the crystal and liquid states, *i.e.* between ordered and disordered states. Hence, the OP seems to identify an aspect of the transition process that is not detected by the Γ^{rv} OP. Consistent with its intent, $KL^{\hat{r}}$ also classifies the alternative structure C_2 as an ordered state. In addition, we note here that we have verified that metadynamics simulations of benzene using $KL^{\hat{r}}$ alone and together with Γ^{rv} lead to the nucleation of multiple ordered states, one of which is the Form I crystal. In these simulations two different kinds of nucleation pathways are sampled. The first pathway is characterized by an initial stage during which the molecules arrange

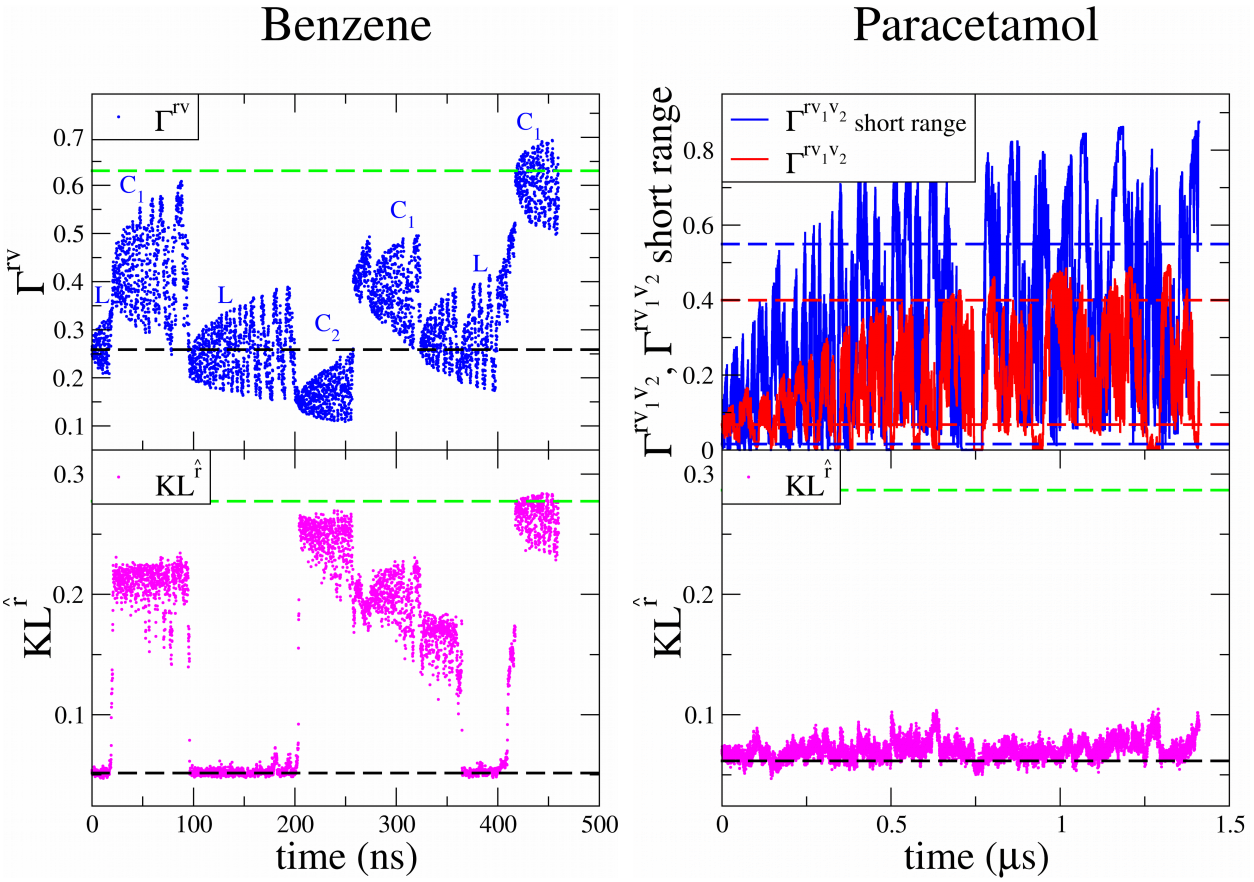

Figure 7: Calculation of the value of $KL^{\hat{r}}$ on the trajectories that were obtained from the metadynamics simulations in which pair-function based OPs were biased. In every panel the typical value of the relative OP in the liquid and crystal state is represented by dashed lines.

approximately in planes and acquire similar orientations. In other words, the distribution of the orientations becomes less random and the system displays orientational ordering. This orientational ordering is then followed by a transition to the Form I crystal. In pathways of the second kind, the positions of the molecular centers become ordered first, *i.e.* the system displays positional ordering. This positional ordering stage is then followed by a transition to the Form I crystal. In contrast, when using only pair-function based OPs to bias the simulations, we only observe nucleation pathways of the first kind.

In the case of paracetamol, the evolution of $KL^{\hat{r}}$ in Fig. 7 shows clearly that during the simulation in which $\Gamma^{rv_1v_2}$ was biased, the system does not visit any ordered phase. In fact, the value of the general positional ordering changed only minimally during the simulation.

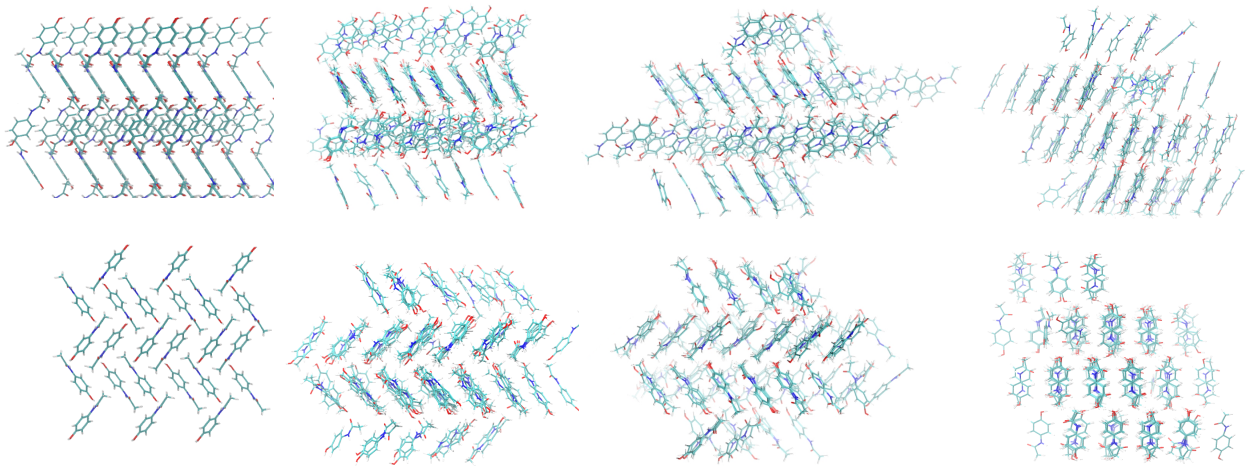

Figure 8: Examples of sampled configurations from the simulation of paracetamol. The first column shows two different views of the reference crystal structure. The other columns show configurations sampled in the metadynamics simulation where $KL^{\hat{r}}$ and $\Gamma^{rv_1v_2}$ were used. The second column shows an example of configuration that is similar to form 1. The value of $\Gamma^{rv_1v_2}$ for this configuration is lower than that of the crystal reference due to the presence of some residual orientational disorder. The third column shows a more ordered form 1 like configuration. It can be noticed however, that part of the crystal is misaligned. The fourth column shows an example of an alternative ordered structure visited by the system during the simulation.

Motivated by these results, we decided to run new metadynamics simulations of paracetamol biasing the $KL^{\hat{r}}$ and $\Gamma^{rv_1v_2}$ OPs. During the simulation time of 1.6 μs, the system visited several ordered states multiple times. Some of these states are very similar to the Form I crystal. Examples of the ordered configurations sampled during the simulation are shown in Fig. 8 and Fig. 9. In Fig. 8, every column shows two different views of the same configuration. The first column shows two different views of the perfect crystal for reference; the second and third columns show two Form I like configurations; and the fourth column shows an example of a different ordered phase.

The configuration shown in the second column has a value of $KL^{\hat{r}}$ that is compatible with that of the reference crystal but the value of $\Gamma^{rv_1v_2}$ is somewhat lower than the reference value due to residual disorder in the mutual orientations between the molecules, which can be seen in the picture. By contrast, the configuration shown in the third column is characterized by a higher degree of orientational order and values for both the OPs that are compatible with

those of the crystal state. In the picture, the centers of the molecules are arranged in rows of horizontal planes. In each of these planes, most of the molecules have orientations that are similar to those of the molecules in the planes that are found in the perfect crystal structure. Although most of the configuration is crystalline, it is evident in the background of the picture that in each of these planes some molecules have orientations that are characteristic of those in the plane above or below. This is an effect often seen in the simulation. It is caused in part by an interplay between the finite size of the system and the periodic boundary conditions. In our simulations, we do not force the system to nucleate in a particular direction, and therefore, the crystal formed does not necessarily fit correctly within the periodic boundary conditions since these are commensurate with the reference crystal. This also explains in part the relative difficulty for the system to get to high values of both the OPs at the same time (See Fig. 9). A difference between the reference crystal structure and both of the configurations shown in the second and third column of Fig. 8 is that the lattice spacing between the molecules is incorrect. A significant consequence of this mismatch is that the corresponding intermolecular hydrogen bond network is not correctly formed. Even considering the influence of finite size effects and periodic boundary conditions, these results suggest that the pair-function based OP $\Gamma^{rv_1v_2}$, even when used in combination with the $KL^{\hat{r}}$ OP, is not adequate to correctly describe the nucleation of paracetamol. Despite these issues, it is clear that the addition of the $KL^{\hat{r}}$ OP greatly improves the sampling and that this OP is able to drive the system to ordered configurations.

3.2 Relative information entropy order parameters for paracetamol Form I nucleation

The purpose of the simulations described in the previous section was to test whether a combination of the $KL^{\hat{r}}$ OP and the pair-function OP could lead to nucleation of the Form I crystal of paracetamol. Ideally, the $KL^{\hat{r}}$ OP would drive the system to a configuration with a high degree of positional order, and the pair-function OP would then drive this ordered

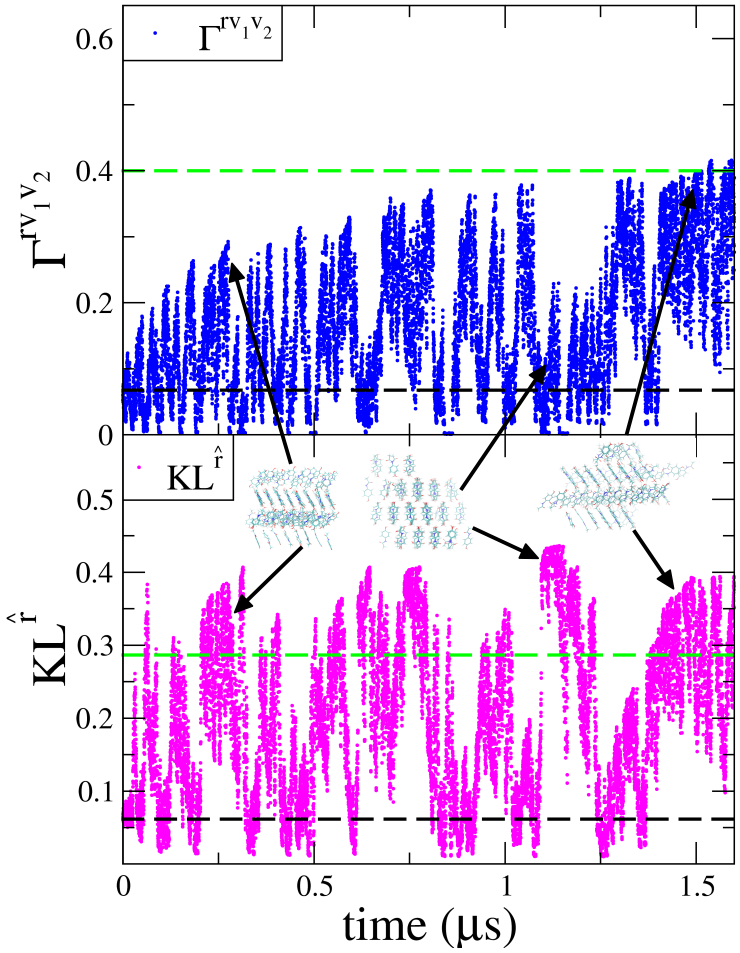

Figure 9: Metadynamics simulation of the paracetamol system obtained biasing both $KL^{\hat{r}}$ and $\Gamma^{rv_1v_2}$. Gaussians of height 6 kJ/mol were deposited every 8 *ps* using a width of 0.007 and 0.01, respectively. The evolution in time of the OPs is shown. Black arrows associate the ordered configurations shown in panel 2, 3, and 4 of Fig. 8 with the portion of simulation during which they were visited.

configuration to the Form I crystal. However, our results demonstrate that biasing the simulation with this combination of OPs is not sufficient to nucleate the Form I crystal. On a more positive note, we can conclude that the increase in configuration space exploration is largely a result of biasing with the $KL^{\hat{r}}$ OP, since previous simulations in which only the pair-function OPs were biased did not sample any ordered configurations. In light of these results, we decided to formulate a more refined set of OPs based solely on the KLD framework. Since our approach is general, any probability distribution of the system can be constructed on the fly, and the corresponding KLD can be utilized to quantify the distance

between this instantaneous distribution and a reference distribution of interest. This is advantageous in multiple ways. First it allows one to build systematically more complex OPs that utilize distributions of additional structural quantities of the system in order to drive nucleation. This is fundamental in a more complex system like paracetamol. A second advantage is that distributions of a target crystal state can be used as reference. In an effort to nucleate the Form I crystal of paracetamol, we have taken advantage of both these possibilities, and we have constructed OPs using distributions of several different structural quantities and compared them to the corresponding distributions of these quantities for the Form I crystal. For the sake of clarity, in what follows, we will denote all the OPs that use a crystal reference distribution with a subscript c.

To detect orientational order of the molecules, we begin by constructing KLD based OPs using the orientation vectors of the molecules, *i.e.* v_1 and v_2. To this end, we first built two separate distributions on the unit sphere of the normalized orientation vectors v_1 and v_2 for all of the molecules in the system. Subsequently, we used the KLD to measure the distance between each of these distributions and their reference distribution in the Form I crystal. We denote these quantities as $KL_c^{v_1}$ and $KL_c^{v_2}$. To avoid the computational cost associated with multidimensional biasing, we combined these two OPs into a single orientational OP that we used to bias our simulations $KL_c^{v_1,v_2} = (KL_c^{v_1} + KL_c^{v_2})/2$. The intuitive idea behind this combination is that it is the simplest way to include the information contained within the two OPs when biasing the system.

In addition to constructing OPs designed to drive the orientational order of the molecules in the system, we also wanted to take the hydrogen bond network into account since it is critical for the stability of the paracetamol crystal and is therefore important for nucleation. Hence, we used OPs to drive the positional ordering of the donor and acceptor atoms involved in hydrogen bonding so that the hydrogen bond network will form naturally. To construct these OPs, we first identified the network of hydrogen bonds present in the Form I crystal. These are shown in Fig. 10 by the dashed lines drawn between $O_a - H_a \cdots O_b$

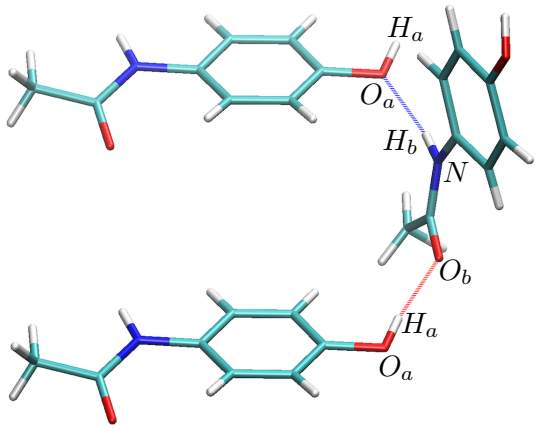

Figure 10: Example of the two unique hydrogen bonds between molecules in the Form 1 crystal.

and $N - H_b \cdots O_a$, respectively. After identifying the hydrogen bonds in the crystal, we built two separate distributions on the unit sphere consisting of the components of each of the normalized distance vectors, denoted $\hat{r}_{OO}$ and $\hat{r}_{ON}$, respectively. The $\hat{r}_{OO}$ distribution was computed from pairwise distance vectors between O_a and O_b atoms, while the $\hat{r}_{ON}$ distribution was computed from pairwise distance vectors between O_a and N atoms. Consistent with the procedure adopted before, we limited ourselves to distance vectors between neighboring molecules. This was achieved using the switching function in eq. (2) with $r_0 = 7$ Å, $n = 10$, and $m = 20$. We then used the KLD to measure the distance between each of these distributions and their reference distribution in the Form I crystal. We denote these quantities $KL_c^{\hat{r}_{OO}}$ and $KL_c^{\hat{r}_{ON}}$, and we combined these two OPs with the $KL_c^{\hat{r}}$ OP, *i.e.* the one comparing the distribution of distances of the centers of the molecule with the reference of Form I. Hence, the positional OP, used for our simulations, becomes $KL_c^{\hat{r},\hat{r}_{OO},\hat{r}_{ON}} = (KL_c^{\hat{r}} + KL_c^{\hat{r}_{OO}} + KL_c^{\hat{r}_{ON}})/3$.

In an attempt to speed up the simulation, we performed a multiple walker metadynamics[15] simulation with ten walkers utilizing the positional and orientational ordering OPs, $KL_c^{\hat{r},\hat{r}_{OO},\hat{r}_{ON}}$ and $KL_c^{v_1,v_2}$ just described. Notice that, because we are using crystal reference distributions in the definition of our OPs and because the KLD is zero if and only if the

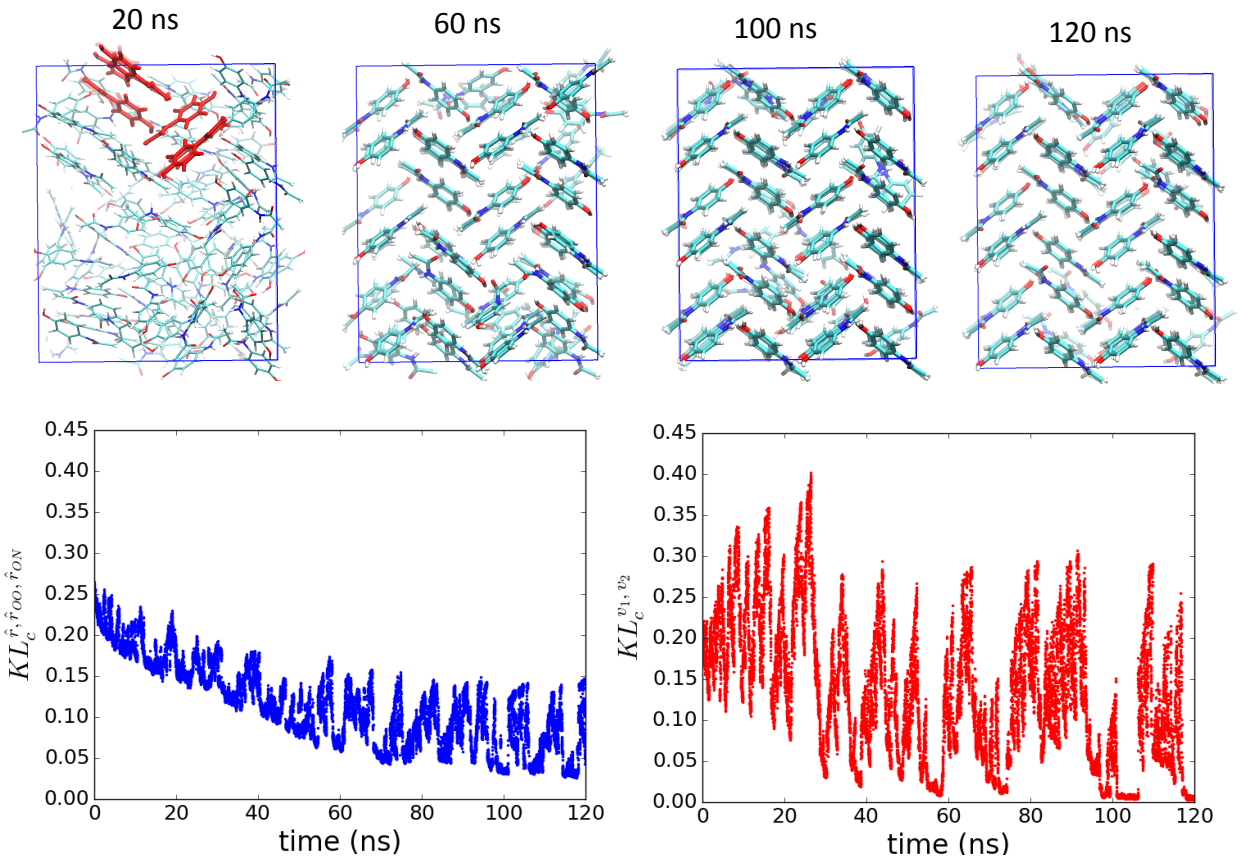

Figure 11: Nucleation trajectory obtained from a multiple walkers metadynamics simulation biasing the positional and orientational OPs, $KL_c^{\hat{r},\hat{r}_{OO},\hat{r}_{ON}}$ and $KL_c^{v_1,v_2}$. The metadynamics bias potential was constructed by deposing Gaussians every 20 ps with a height of 5 kJ/mol with σ values of 0.015 and 0.02 for the $KL_c^{\hat{r},\hat{r}_{OO},\hat{r}_{ON}}$ and $KL_c^{v_1,v_2}$ OPs, respectively. The configurations in the upper panel are labeled by the time in which they were sampled from the trajectory. In the configuration sampled at 20 ns, there is a small nucleus of dimers highlighted in red. This configuration represents the first point along the trajectory where a portion of the crystal has visibly nucleated.

reference distribution is identical to the one computed on the fly, these OPs have low values when the system has a structure that is similar to the Form I crystal. The simulation was performed for a total of 120 ns using the Parrinello-Rahman barostat at a temperature and pressure of 298 K and 1 atm. The results from our multiple walker metadynamics simulations are presented in Fig. 11, where the evolution of the OPs during a nucleating trajectory as well as sampled configurations along the trajectory are shown. Altogether, two of the walkers out of ten nucleated in a 120 ns time frame. Fig. 11 shows that the nucleation event begins around 20 ns where it is clear in the configuration that a small nucleus of dimers have

formed within the system. This nucleus is highlighted in red in the figure. It is interesting to note that the nucleation begins with a formation of dimers. These correspond to the first peak (near 4 Å) in the joint distribution in Fig. 3 and represent the basic subunit of the Form I crystal. In addition, we observe that after this dimer subunit has formed, the system rapidly orders, which is evident by the nearly fully formed crystal present in the sampled configuration at 60 ns. Also evident from Fig. 11 is the difference in the behavior and range of fluctuations in the time series of the positional and orientational OPs. The positional OP shows an overall downward trend, while the orientational OP fluctuates between two basins. The first of these basins is predominant during the first 30 ns and the value of the OP is higher when the system is inside this basin suggesting that the structure is more disordered. The second basin, for which the OP has a lower value, dominates during the later parts of the simulation. The wide range of fluctuations of the orientational OP is at least partially due to the fact that the distributions of orientation vectors are inherently more noisy as fewer data points are used to construct the distributions that measure the orientational order. Only one orientation vector per molecule is used to construct the distributions for the orientational OP whereas multiple bond vectors per molecule are used to construct the distributions for the positional OP.

The very slow evolution of the positional OP towards low values suggests that this OP identifies a key bottleneck in the nucleation of the system. After analyzing the time series of the three OPs that comprise the positional OP, we observed that the $KL_c^{\hat{r}_{OO}}$ and $KL_c^{\hat{r}_{ON}}$ OPs are the slowest evolving OPs. Given that these OPs drive the positional ordering of the donor and acceptor atoms involved in hydrogen bonding, this suggests that formation of the correct hydrogen bond network is the rate-limiting step in the nucleation. In addition, hydrogen bonding seems to play an active role in the formation of the nucleus. We noticed numerous times throughout a nucleation trajectory that once one of the hydrogen bonds formed between adjacent paracetamol molecules in the correct direction and orientation, the hydrogen bond seemed to anchor the paracetamol molecules together, stabilizing the forming

complex for a period of time before random fluctuations allowed it to overcome the large rotational energy barriers present in the system (and incorporate into the crystal lattice). Given the importance of the hydrogen bond network, inclusion of the $KL_c^{\hat{r}_{OO}}$ and $KL_c^{\hat{r}_{ON}}$ OPs in the positional OP is a necessary component to drive nucleation in the system. However, despite the fact that hydrogen bonding is very important for nucleation, we observed that the dimer subunit is the first part of the crystal that is visibly evident during nucleation, and interestingly enough, there are no hydrogen bonds present between adjacent molecules in the dimer subunit. It is unclear whether or not this dimer subunit is stabilized through van der Waals interactions or by hydrogen bonds from surrounding molecules. Consequently, the nucleation mechanism is more complex than would appear and requires a further detailed study to get a complete molecular level understanding.

4 Conclusions

In this paper, we have introduced a novel and general approach for the construction of OPs that are capable of driving the nucleation of molecular crystals. Our approach works by constructing reduced dimensional distributions of relevant quantities of the system and computing the relative information entropy between these distributions and reference distributions. We have shown how it is possible to construct OPs that characterize order in a system without reference to a specific crystal form. In this framework, once essential structural quantities are identified, the system is considered to be ordered when the distributions of these quantities become peaked, *i.e.* when there is a reduction in the system entropy. OPs of this kind are particularly interesting as they do not make specific assumptions about the mechanism of nucleation. In particular, considering the specific case of nucleation from the melt of 144 benzene molecules, we have examined how an OP of this kind distinguishes and correctly classifies ordered and disordered states, highlighting aspects of the nucleation process that remained hidden from the pair-function based OPs that have been previously used

to study this kind of transition. We have also shown that in the case of paracetamol, the approach that relies solely on pair-functions based OPs fails completely. On the contrary, the addition of one OP based on our approach, and aimed at detecting positional order, dramatically increases the exploration of the phase space and allows the system to visit ordered states. However, it does not induce nucleation to crystal Form I. Finally, thanks to the flexibility of our approach, we were also able to target the nucleation of the Form I crystal by constructing OPs that used specific distributions of the crystal as a reference state. These distributions were constructed from specific molecular orientations, center of mass positions, and the positions of donor and acceptor atoms involved in hydrogen bonding. Inclusion of the latter of these quantities was crucial for the nucleation.

Our approach significantly enlarges the spectrum and quality of OPs that can be formulated to study the nucleation of molecular crystals. Moreover, using relative information entropy between distributions to formulate descriptors for the study of nucleation is philosophically appealing since what characterizes the transition of a system from a disordered to an ordered phase is the emergence of long range correlations and the decrease of entropy. Our approach is also versatile because distributions of any set of structural quantities can be considered and different levels of description can be combined until the description is refined enough for the transition to be elucidated. As our simulations of paracetamol show, the possibility of formulating different, complementary levels of description is a very important and often necessary feature to guide the construction of OPs that are able to drive the nucleation process in challenging cases.

The approach we have introduced is intrinsically global as the distributions we compare are built using structural quantities that are derived from all the molecules in the system. However, to study nucleation in larger systems it may be necessary to track the degree of order in the system using an OP that has a more local resolution. One way to achieve this, would be to limit the calculation of the distribution to a subset of the molecules that occupy a particular, limited portion of space. It would also be interesting to try to develop a per-

molecule version of these OPs using an approach similar to that of Piaggi and Parrinello.[49] Finally, we also note that even when pair-function based OPs fail to drive nucleation in biased simulations, they can still be used as analysis tool in post processing.

Methods

All simulations discussed in this paper have been run using PLUMED 2[50] with GROMACS[51] (version 4.6.7 for the simulations described in section 2 and version 5.1.4 for all the others) using the CHARMM36 CGenFF force field (the point charge distribution was reparametrized in the case of paracetamol[41]). All simulations were carried out using a time step of 2 fs and long range electrostatic interactions were treated with the particle mesh Ewald approach. All bonds were constrained using the SHAKE[52] algorithm and the Nose-Hoover thermostat[53,54] was used to thermalize the system at 250 K and 298 K for benzene and paracetamol, respectively. The simulation box was orthorhombic in the case of benzene and monoclinic for paracetamol. Crystal structures were downloaded from the CCSD database[55] and equilibrated for 10 ns using the Parrinello-Rahman barostat.[56] The box size was then kept fixed and all other simulations were run in NVT, unless otherwise noted. To generate initial configurations for MD simulations in the liquid state we used PACKMOL[57] and filled the simulation box with 144 benzene molecules in one case and with 96 paracetamol ones in the other. The resulting configurations were then relaxed in NVT for 50 ns.

Appendix A Pair-function based order parameters

Santiso and Trout[34] introduced a systematic method for developing pair-distribution function based OPs. Starting from the point molecule representation, one builds an OP by modeling a pair-distribution density function based on distances and orientations and parameterizing it to a specific crystal form. This parametrization is done by performing a MD simulation of the reference crystal form and by fitting the pair-function to the distribution generated

by MD up to a given cutoff distance from a molecule. The model pair-function used takes the generic form

$$F\left(x_k, x_l\right) = \sum_{\alpha=1}^{M} f_\alpha\left(x_k, x_l\right) = \sum_{\alpha=1}^{M} \prod_{\beta\in\mathcal{B}} f_\alpha^\beta\left(x_k^\beta, x_l^\beta\right), \tag{7}$$

where x_k and x_l denote the general point molecule representation of the k-th and l-th molecules, α loops over the M peaks within the selected distance cutoff in the pair-distribution function, and f_α is chosen to be a product of independent distributions over the set $\mathcal{B}$ of attributes of the point representation that are being considered. A global OP is then defined by summing over all pairs of molecules lying within the range considered, while a local OP characterizing the order around any given molecule i can be defined by summing over all other molecules j within the cutoff. The explicit form of such a local OP is

$$G_i^{\mathcal{B}} = {\sum_{j\neq i}}' F\left(x_i, x_j\right) = {\sum_{j\neq i}}' \sum_{\alpha=1}^{M} \prod_{\beta\in\mathcal{B}} f_\alpha^\beta\left(x_i^\beta, x_j^\beta\right), \tag{8}$$

where the prime over the sum indicates the distance cutoff restriction. This allows one to define a family of per-molecule OPs by selecting only certain attributes of the point molecule representation. For instance, if only the modulus of the distance between molecule i and its neighbors is used one obtains a distance order parameter[34]

$$G_i^r = {\sum_{j\neq i}}' \sum_{\alpha=1}^{M} f_\alpha^r(|r_i - r_j|), \tag{9}$$

where r_i and r_j are the Cartesian coordinates of the centers of the point molecule representation of molecule i and j respectively, and f_α^r can be modeled using a Gaussian function. A more refined description can be obtained by adding relative orientations between the vectors of the point representation

$$G_i^{rv} = {\sum_{j\neq i}}' \sum_{\alpha=1}^{M} f_\alpha^r(|r_i - r_j|) f_\alpha^v(v_i, v_j), \tag{10}$$

where v_i and v_j are the vectors (easily extendable to more than one for each molecule) defining the orientation of molecules i and j and f_α^v can be parametrized, for example, by a von Mises distribution.[58] Other layers of description can be obtained by using bond orientations or any internal degree of freedom, if present.[34] Summing over the molecules belonging to the different cells of a spatial partition of the system, one can have a set of spatially localized OPs. This approach has been used in combination with the string method to study the nucleation of benzene from the melt[28,29] and other systems.[59]

Salvalaglio and co-workers[35] have used a similar approach but, instead of trying to accurately reproduce the pair-distribution function, they opted to use a simple unnormalized Gaussian for every peak. Moreover, they introduced a smooth switching function, s, that acts on the intermolecular distances and selects the pairs of molecules within the distance cutoff. Following the spirit of their formulation [‡] one can obtain an OP based only on relative orientations:

$$\Gamma_i^v = \frac{1}{n_i} \sum_{j \neq i} s(|r_i - r_j|) \sum_{\alpha=1}^{M} e^{-((\theta(v_i,v_j)-\theta_\alpha)^2/2\sigma_{\theta_\alpha}^2)}. \tag{11}$$

Extending this formulation to include the definition of peaks in a multidimensional space, leads naturally to the OPs defined by equations (1) and (3).

To better illustrate how these OPs can discriminate between different phases, Fig. 12 shows the average distribution over all the molecules of the values of the per-molecule OP, Γ_i^{rv} of eq. (1), for benzene molecules in the liquid (red) and crystal (blue) phase. It is evident from Fig. 12 that the two distributions are well separated.

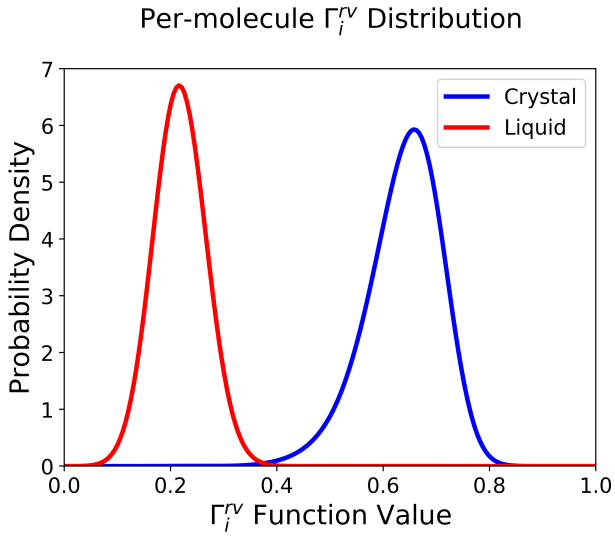

Figure 12: Average distribution over all the molecules of the values of the per-molecule OP, Γ_i^{rv}, for the liquid and the crystal states. The distribution for the liquid phase and crystal phase are separated. This separation is the minimal requirement for a function to be a good OP to describe the transition between the two phases.

Appendix B Degeneracies in pair-function based order parameters

The simulations discussed in section 2 show a number of potential issues associated with the use of pair-function based OPs. In Fig. 13, we highlight some of the issues with these OPs that we observed from our metadynamics simulations. In Fig. 13 every panel shows (in blue) the probability density in distances and angles space for a chosen configuration of either the benzene or the paracetamol system. For comparison, these density distributions are overlaid with the reference density for the Form I crystal state, which is shown with black contour lines. The top panel shows the probability distribution for a configuration of benzene whose value of Γ^{rv} is large but which is seen to be disordered upon visual inspection. The density distribution for this configuration is concentrated in areas overlapping with the

[‡]The formulation of Salvalaglio and co-workers also includes a prefactor that acts as a switching function of the coordination number that allows the OP to be effective even in the case of nucleation from solution.[32,33] This prefactor is irrelevant when considering nucleation from the melt, and we omit it throughout the rest of the presentation.

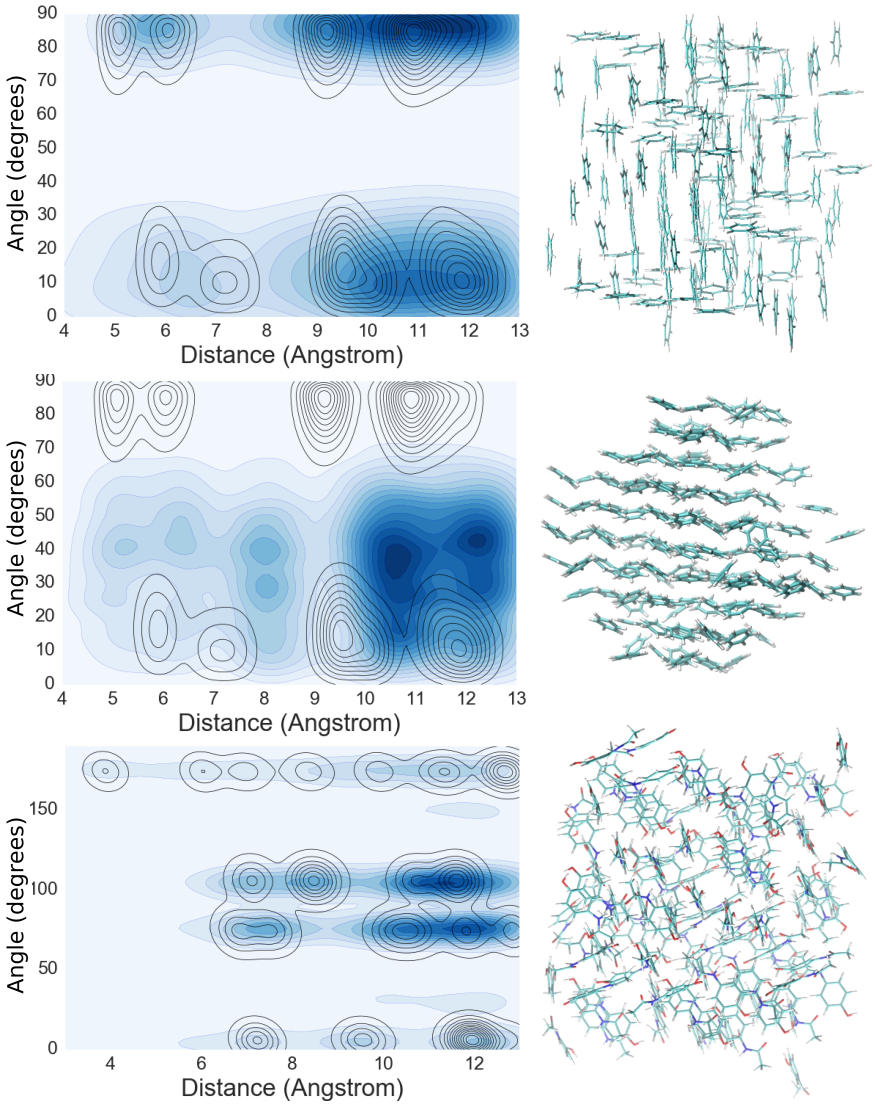

Figure 13: Examples of states that are misclassified by pair-functions based OPs. In every panel the instantaneous joint distribution of distances and relative angles is computed for a selected configuration of the system and is shown in shades of blue. Superimposed black contour lines represent the reference distribution computed for the crystal state. Next to every probability density, a projected view of the corresponding configuration is shown. The first panel shows an example of a disordered benzene configuration for which the corresponding value of Γ^{rv} is high. Even though it is disordered, this configuration has a high value for the CV because its density is concentrated around the peaks rewarded by Γ^{rv}. The second panel shows the joint density for a configuration of benzene in the C_2 crystal state. None of peaks in the instantaneous distribution overlap with those typical of Form I. The third panel shows the joint density for distances and angles between the first vectors v_1 for a paracetamol configuration whose value of $\Gamma^{rv_1v_2}$ is high. Even though the configuration is amorphous its density is concentrated around the peaks typical of the distribution of Form I.

peaks characterizing the Form I crystal but not all of these are correctly populated. This phenomenon is often seen in our metadynamics simulations. It happens frequently because there are enumerable ways to populate the peaks of the reference distribution that all give a large value for the OP. In other words, there are a number of configurations that have degenerate values of the OP, but only one of these configurations actually corresponds to the Form I crystal. The middle panel shows the density for a configuration of benzene in the alternative crystal form we have labeled C_2. Since the peaks in the joint distribution characterizing this metastable state have a minimal overlap with those of the Form I reference distribution, it is clear why configurations in this state have values of Γ^{rv} that are comparable and even smaller than those observed for the liquid state. The bottom panel shows the joint distribution of distances and angles between the first vector v_1 for a configuration of paracetamol characterized by a large value of $\Gamma^{rv_1v_2}$ but which is nowhere near a crystal phase. Also in this case, the distribution is concentrated around some of the peaks characterizing Form I but most of the peaks are not substantially populated.

Acknowledgements

We gratefully acknowledge support from the MIT-Novartis Center for Continuous Manufacturing.

References

(1) Hartel, R. W. *Crystallization in Foods*; Aspen Publishers: Gaithersburg, MD, 2001.

(2) Lee, A. Y.; Myerson, A. S. Particle engineering: Fundamentals of particle formation and crystal growth. *MRS Bull.* **2006**, *31*, 881–886.

(3) Paul, E. L.; Tung, H.-H.; Midler, M. Organic crystallization processes. *Powder Technol.* **2005**, *150*, 133–143.

(4) Shekunov, B. Y.; York, P. Crystallization processes in pharmaceutical technology and drug delivery design. *J. Cryst. Growth* **2000**, *211*, 122–136.

(5) Gasser, U.; Weeks, E. R.; Schofield, A.; Pusey, P.; Weitz, D. Real-space imaging of nucleation and growth in colloidal crystallization. *Science* **2001**, *292*, 258–262.

(6) Harano, K.; Homma, T.; Niimi, Y.; Koshino, M.; Suenaga, K.; Leibler, L.; Nakamura, E. Heterogeneous nucleation of organic crystals mediated by single-molecule templates. *Nat. Mater.* **2012**, *11*, 877–881.

(7) Schreiber, R. E.; Houben, L.; Wolf, S. G.; Leitus, G.; Lang, Z.-L.; Carbó, J. J.; Poblet, J. M.; Neumann, R. Real-time molecular scale observation of crystal formation. *Nat. Chem.* **2017**, *9*, 369–373.

(8) Carter, E.; Ciccotti, G.; Hynes, J.; Kapral, R. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.* **1989**, *156*, 472–477.

(9) Sprik, M.; Ciccotti, G. Free energy from constrained molecular dynamics. *J. Chem. Phys.* **1998**, *109*, 7737.

(10) Marinari, E.; Parisi, G. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **1992**, *19*, 451.

(11) Hansmann, U. H. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

(12) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(13) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **1998**, *108*, 1964–1977.

(14) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562.

(15) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* **2006**, *110*, 3533–3539.

(16) Wang, F.; Landau, D. P. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.

(17) Rosso, L.; Mináry, P.; Zhu, Z.; Tuckerman, M. E. On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *J. Chem. Phys.* **2002**, *116*, 4389.

(18) Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **2006**, *426*, 168–175.

(19) Abrams, J. B.; Tuckerman, M. E. Efficient and direct generation of multidimensional free energy surfaces via adiabatic dynamics without coordinate transformations. *J. Phys. Chem. B* **2008**, *112*, 15742–15757.

(20) Ciccotti, G.; Meloni, S. Temperature accelerated Monte Carlo (TAMC): a method for sampling the free energy surface of non-analytical collective variables. *Phys. Chem. Chem. Phys.* **2011**, *13*, 5952–5959.

(21) Van Erp, T.; Moroni, D.; Bolhuis, P. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* **2003**, *118*, 7762.

(22) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125*, 024106.

(23) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **2008**, *128*, 144120.

(24) Allen, R.; Valeriani, C.; Rein ten Wolde, P. Forward flux sampling for rare event simulations. *J. Phys.-Condens. Mat.* **2009**, *21*, 463102.

(25) Gobbo, G.; Laio, A.; Maleki, A.; Baroni, S. Absolute Transition Rates for Rare Events from Dynamical Decoupling of Reaction Variables. *Phys. Rev. Lett.* **2012**, *109*, 150601.

(26) Gobbo, G.; Leimkuhler, B. J. Extended Hamiltonian approach to continuous tempering. *Phys. Rev. E* **2015**, *91*, 061301.

(27) Shah, M.; Santiso, E. E.; Trout, B. L. Computer simulations of homogeneous nucleation of benzene from the melt. *J. Phys. Chem. B* **2011**, *115*, 10400–10412.

(28) Bellucci, M. A.; Trout, B. L. Bezier curve string method for the study of rare events in complex chemical systems. *J. Chem. Phys.* **2014**, *141*, 074110.

(29) Santiso, E. E.; Trout, B. L. A general method for molecular modeling of nucleation from the melt. *J. Chem. Phys.* **2015**, *143*, 174109.

(30) Giberti, F.; Salvalaglio, M.; Mazzotti, M.; Parrinello, M. Insight into the nucleation of urea crystals from the melt. *Chem. Eng. Sci.* **2015**, *121*, 51–59.

(31) Pietrucci, F.; Martoňák, R. Systematic comparison of crystalline and amorphous phases: Charting the landscape of water structures and transformations. *J. Chem. Phys.* **2015**, *142*, 104704.

(32) Salvalaglio, M.; Mazzotti, M.; Parrinello, M. Urea homogeneous nucleation mechanism is solvent dependent. *Faraday Discuss.* **2015**, *179*, 291–307.

(33) Salvalaglio, M.; Perego, C.; Giberti, F.; Mazzotti, M.; Parrinello, M. Molecular-dynamics simulations of urea nucleation from aqueous solution. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E6–E14.

(34) Santiso, E. E.; Trout, B. L. A general set of order parameters for molecular crystals. *J. Chem. Phys.* **2011**, *134*, 064109.

(35) Salvalaglio, M.; Vetter, T.; Giberti, F.; Mazzotti, M.; Parrinello, M. Uncovering molecular details of urea crystal growth in the presence of additives. *J. Am. Chem. Soc.* **2012**, *134*, 17221–17233.

(36) Diao, Y.; Myerson, A. S.; Hatton, T. A.; Trout, B. L. Surface design for controlled crystallization: The role of surface chemistry and nanoscale pores in heterogeneous nucleation. *Langmuir* **2011**, *27*, 5324–5334.

(37) Shtukenberg, A. G.; Lee, S. S.; Kahr, B.; Ward, M. D. Manipulating crystallization with molecular additives. *Annu. Rev. Chem. Biomol.* **2014**, *5*, 77–96.

(38) Tan, L.; Davis, R. M.; Myerson, A. S.; Trout, B. L. Control of heterogeneous nucleation via rationally designed biocompatible polymer surfaces with nanoscale features. *Cryst. Growth Des.* **2015**, *15*, 2176–2186.

(39) Ward, M. D. Soft Crystals in Flatland: Unraveling Epitaxial Growth. *ACS Nano* **2016**, *10*, 6424–6428.

(40) Frank, D. S.; Matzger, A. J. Influence of Chemical Functionality on the Rate of Polymer-Induced Heteronucleation. *Cryst. Growth Des.* **2017**, *17*, 4056–4059.

(41) Stojakovic, J.; Baftizadeh, F.; Bellucci, M. A.; Myerson, A. S.; Trout, B. L. Angle-directed nucleation of paracetamol on biocompatible nano-imprinted polymers. *Cryst. Growth Des.* **2017**,

(42) Wijethunga, T. K.; Baftizadeh, F.; Stojakovic, J.; Myerson, A. S.; Trout, B. L. Experimental and mechanistic study of the heterogeneous nucleation and epitaxy of acetaminophen with biocompatible crystalline substrates. *Cryst. Growth Des.* **2017**,

(43) Kullback, S. *Information Theory and Statistics*; John Wiley: New York, 1959.

(44) Silverman, B. W. *Density Estimation for Statistics and Data Analysis*; Chapman & Hall/CRC press: Boca Raton, FL, 1986; Vol. 26.

(45) Lin, J. Divergence measures based on the Shannon entropy. *IEEE T. Infor. Theory* **1991**, *37*, 145–151.

(46) Jadrich, R.; Lindquist, B.; Truskett, T. Probabilistic inverse design for self-assembling materials. *J. Chem. Phys.* **2017**, *146*, 184103.

(47) Gimondi, I.; Salvalaglio, M. CO2 packing polymorphism under confinement in cylindrical nanopores. 2017, arXiv:1709.04586. arXiv.org e-Print archive. https://arxiv.org/abs/1709.04586 (accessed Nov. 30, 2017)

(48) Piaggi, P. M.; Valsson, O.; Parrinello, M. Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations. *Phys. Rev. Lett.* **2017**, *119*, 015701.

(49) Piaggi, P. M.; Parrinello, M. Entropy based fingerprint for local crystalline order. *J. Chem. Phys.* **2017**, *147*, 114112.

(50) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(51) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(52) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Chem. Phys.* **1977**, *23*, 327–341.

(53) Nosé, S. A unified formulation of the constant temperature molecular dynamics methodsJ. *Chem. Phys.* **1984**, *81*, 511–519.

(54) Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phy. Rev. A* **1985**, *31*, 1695.

(55) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B* **2002**, *58*, 380–388.

(56) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(57) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Computat. Chem.* **2009**, *30*, 2157–2164.

(58) Mardia, K. V.; Jupp, P. E. *Directional Statistics*; John Wiley & Sons: Chichester, West Sussex, United Kingdom, 2000.

(59) He, X.; Shen, Y.; Hung, F. R.; Santiso, E. E. Molecular simulation of homogeneous nucleation of crystals of an ionic liquid from the melt. *J. Chem. Phys.* **2015**, *143*, 124506.

For Table of Contents Use Only

Nucleation of molecular crystals driven by relative information entropy

Authors: Gobbo Gianpaolo, Michael A. Bellucci, Gareth A. Tribello, Giovanni Ciccotti, Bernhardt L. Trout

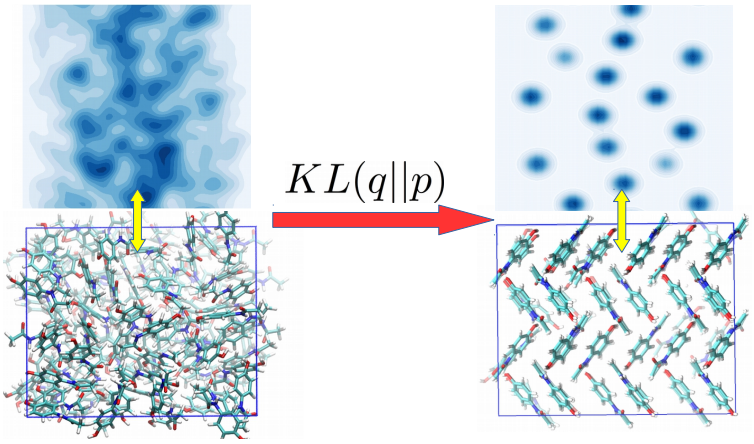

Figure 14: Table of Contents Figure.